Ciência Téc. Vitiv. 15 (1), 15-26. 2000

STATISTICAL METHODS FOR CONTROLLING WINE TASTING PANELS

MÉTODOS ESTATÍSTICOS PARA O CONTROLO DE CÂMARAS DE PROVADORES DE VINHOS

Helena Alvelos[1], J. A. Cabral[2], B. Amaral[3] P. Barros[3]

[1]Secção Autónoma de Gestão e Engenharia Industrial, Universidade de Aveiro, 3810 Aveiro, Portugal ; e.mail: helena@egi.ua.pt
[2]Faculdade de Engenharia da Universidade do Porto, Rua dos Bragas, 4099 Porto Codex, Portugal; e.mail: jacabral@fe.up.pt
[3]Instituto do Vinho do Porto, 4050 Porto, Portugal; e.mail: bamaral@mail.ivp.pt

(Manuscrito recebido em 02.03.00. Aceite para publicação em 31.05.00)

SUMMARY

Sensory analysis is an essential tool in the wine evaluation. Although it has a long tradition, scientific methods for the control of sensory analysis are scarce and often not practical. This paper presents some new procedures that were developed during the accreditation process of the Tasting Panel of the Port Wine Institute.

Key-words: statistical process control, CUSUM tabulation charts, logistic regression, sensory analysis

Palavras chave: controlo estatístico de processos, cartas de controlo CUSUM, regressão logística, análise sensorial

INTRODUCTION

The sensory evaluation of wines is commonly performed by tasting panels. Like any other measurement instrument, tasting panels need to be calibrated and controlled. Many authors present methods that are used to compare the tasters' performance (see, for example, Lima *et al.*, 1988; Mangan, 1992; Sinesio *et al.*, 1990; Cardinal *et al.*, 1994), but information on methods for the systematic quality assessment of tasting panels' decisions is scarce and often not practical.

During the accreditation process, The Port Wine Institute (IVP) was faced

with the lack of sound and proved methodologies for the assessment and control of the sensory evaluation of wines. This fact induced IVP in developing new procedures for the statistical control of the Tasting Panel (TP).

Whenever a company intends to commercialise a new batch of a particular type of Port wine, a sample must be submitted to the TP. Based on the organoleptic characteristics of the sample the tasting panel has to judge if the wine is approved or not. The final decision ("accepted" or "not accepted") depends on the individual judgements of each panel member: if the panellists majority do not approve the sample, the wine is rejected, otherwise the wine is accepted. This situation prevails in many other institutions that are responsible for supervising and controlling the wines denomination of origin.

Two topics are addressed in this paper: the control of the stability of each panellist decision process over time and the assessment of the reliability of panellists perceptions. Based on the results of the referred assessment, a new procedure for establishing the tasting panel final decision is proposed.

CONTROLLING THE STABILITY OF THE INDIVIDUAL DECISION PROCESS OVER TIME

The procedures presented in this section assume that during the tasting session each panellist evaluates different (at least two) sensory attributes of the wine sample (for example, *colour*, *aroma*, *taste*, *body*) using a quantitative scale or an ordinal scale with at least five levels (for instance, 1 – "very poor"; 2 – "poor", 3 – "acceptable", 4 – "good", 5 – "very good"). It is also presumed that, based on the evaluation of those sensory characteristics, the panellist assesses the overall quality of the sample and decides whether or not the wine shall be approved. The decision is a binary variable that can only take the values of "zero" - if the wine is rejected - or "one" - if the wine is accepted. A discussion about how to organise a tasting evaluation form for Port wines can be found in Van Zeller (1984).

If two different samples receive the same marks on the characteristics under evaluation it is expected that the final decision of the panellist will also be the same for both samples. Otherwise, it can be concluded that the panellist has not a structured and stable decision process over time. Two aspects of this problem are addressed: how to model the individual decision process and how to monitor the stability of this process over time.

MODELLING THE INDIVIDUAL DECISION PROCESS

The decision process of each panellist is modelled through a logistic regression (Freund and Wilson, 1998; Agresti, 1990), where the independent variables

(X_1, X_2, ... X_i) are the sensory or organoleptic characteristics under evaluation and the dependent variable (Y) represents the panellist final decision ("not accepted" or "accepted", 0 or 1). This model, that takes into account the binary nature of the dependent variable, is defined by:

$$Y = \frac{e^{\alpha+\beta_1 \cdot X_1+\beta_2 \cdot X_2+\beta_3 \cdot X_3+\cdots+\beta_i \cdot X_i}}{1+e^{\alpha+\beta_1 \cdot X_1+\beta_2 \cdot X_2+\beta_3 \cdot X_3+\cdots+\beta_i \cdot X_i}} \quad (1)$$

The model parameters are estimated using the maximum likelihood procedure (see Hosmer and Lemeshow, 1989) that is available in many commercial statistical packages.

In order to obtain good estimates for the model parameters (a, b_1, b_2, ... , b_i) it is recommended that the calibration data (from which the parameters are derived) includes at least 300 samples. In the IVP, this amount of data corresponds for each panellist to a period of about 4 months. The data should be previously "cleaned" of outliers and suspicious values (see Belsley, 1991).

The cases classification table, also available in many statistical packages, is developed using the logistic regression model and the calibration data. This table, which shows the goodness of fit of the model to the actual decision process of the panellist, is obtained by counting the number of times that the model correctly predicted the actual panellist decisions.

Table I presents an example of two cases classification tables obtained in a real situation with two panellists (A and B), using a logistic regression model with four independent variables (X_1, X_2, X_3 and X_4).

Taking for example panellist A, it can be seen that the model predicted that the decision should be 0 ("not accepted") 74 times (69 + 5) out of 328 samples tasted. However, the true number of "not accepted" results is 69 + 2. In fact, panellist A did not approve two samples for which the model predicted the result 1 ("accepted").

On the other hand, comparing the coefficients of both models, it can be concluded that panellist B gives more importance to the characteristic represented by the variable X_4 than to the characteristic represented by X_3. On the contrary, panellist A is more sensible to X_3 than to X_4.

MONITORING THE INDIVIDUAL DECISION PROCESS OVER TIME

Assuming that the logistic model correctly describes the normal behaviour of the panellist, the cases classification tables can help in monitoring the panellist decision process over time. For example, taking the "not accepted" decisions of panellist A, it is expected that the agreement between the model and his

actual opinions will occur with a frequency of 97.2% (see Table I). In other words, the expected disagreement, p_R, is 2.8%. In the same manner, the expected disagreement concerning the "accepted" decisions is p_A = 1.9%.

TABLE I

Classification of cases
Classificação dos casos

	Panellist A			**Panellist B**		
	Predicted: 0	*Predicted: 1*	*Agreement [%]*	*Predicted: 0*	*Predicted: 1*	*Agreement [%]*
Decision: 0	69	2	97.2%	45	8	84.9%
Decision: 1	5	252	98.1%	9	280	96.9%
(a = -20.2; b_1 = 4.1; b_2 = 1.9; b_3 = 1.1; b_4 = 0.6)				(a = -32.0; b_1 = 8.6; b_2 = 3.0; b_3 = 1.1; b_4 = 1.7)		

For each type of decisions ("accepted", for instance), the number of disagreements (y) can be interpreted as a binomial variable (assuming that the decisions are independent) with parameters N_A - the actual number of accepted wines - and p_A - the disagreement proportion for the "accepted" decisions.

In order to monitor the stability of the individual decision process over time, the two null hypothesis $p_{A0} = p_A$ and $p_{R0} = p_R$ must be tested periodically. Remember that p_A and p_R represent the "disagreement" values obtained with the logistic regression model over the calibration period for, respectively, the "accepted" and "not accepted" decisions. Each individual decision process is monitored using two Cumulative-Sum (CUSUM) tabulation charts (Harrison *et al.*, 1986), one for each type of decisions: "approved" and "not approved". The CUSUM statistic (y_A or y_R) is the number of disagreements observed in a sample of size N_A or N_R (N_A being the actual number of wines approved and N_R the actual number of wines not approved). The tabulation procedure is preferred over the V-mask scheme, due to its easier computation.

The CUSUM parameters are selected following the procedures recommended by the British Standard BS 5703 and presented in Harrison *et al.*, 1986. Two alternative schemes – C1 and C2 – are feasible. The C2 schemes are preferred, since they are faster than C1 in detecting a change in p (p being p_A or p_R). The corresponding false-alarm rate (α) is also higher in C2 than in C1 schemes (between 0.4% and 0.5%). In practical terms, this range of values is acceptable.

Appendix A presents the tables (Table AI and Table AII) used to select the parameters for the CUSUM tabulation under scheme C2. If p_A or p_R are lower than 0.1 ($p < 0.1$) the binomial distribution can be approximated by a

Poisson distribution and the CUSUM schemes for the number of nonconformities are base on Table AI values. Otherwise, if $p > 0.1$, the binomial distribution is considered and the values for the parameters are presented in Table AI: CUSUM schemes for number of nonconforming units. In practice, when developing a CUSUM tabulation the following procedure should be used:

(i) Calculate $c = N \cdot p$ (N being N_A or N_R)

(ii) If $p < 0.1$ obtain the "Decision Interval" H and the parameter F from Table AI. If $p > 0.1$ obtain H and F from Table AII. If p and c values are not tabulated, linear interpolation in both H and F is recommended (Harrison et al., 1986).

(iii) Calculate the reference values $K_1 = c + F$ and $K_2 = c - F$.

(iv) For each sample j compute the CUSUM values $C^1_j = C^1_{j-1} + (y_j - K_1)$ (to detect an upper trend in p) and $C^2_j = C^2_{j-1} + (y_j - K_2)$ (to detect a lower trend in p). The values of C^1_0 and C^2_0 are equal to 0.

(v) If $C^1_j > H$ then the chart emits an upper signal. If $C^1_j < 0$ reset the chart making $C^1_j = 0$.

(vi) If $C^2_j < -H$ then the chart emits a lower signal. If $C^2_j > 0$ reset the chart making $C^2_j = 0$.

An out-of-control point signal is considered a strong evidence that the panellist is no longer using the same decision process. The model should be recalibrated periodically or when it is suspected that the panellist changed his behaviour significantly.

When the expected proportion of disagreements is too large (say, greater than 20%) or too small (for instance lower then 0.5%) it is not recommended the use of control charts. In the first case, it is clear that the logistic regression is not modelling the panellist decision process adequately. It should be investigated why. In the latter case, the adjustment is so perfect that whenever it a single case of misclassification occurs the situation must be analysed (the use of control charts is, then, unnecessary).

The next sub-section presents an example of the use of the CUSUM tabulation in monitoring the decision process of panellist B ("not approved" decisions) presented in Table I. The figures concerning the number of disagreements between the model and the actual decision of panellist B were simulated.

AN EXAMPLE OF A CUSUM TABULATION IN MONITORING A DECISION PROCESS

The simulation refers to the "not approved" decisions produced by panellist B (see table I) over 6 samples including 50 wines each. In this situation $p_R = 1 - 0.849 = 0.151$ and $N_R = 50$. The number of disagreements between the

logistic model and panellist B "not approved" decisions is presented in the second column of Table II.

TABLE II

CUSUM tabulation on "not approved" decisions of panellist B

Tabela CUSUM para as decisões "não aprovado" do provador B

Sample number (j)	*Number of disagreements* (y_{Rj})	C^1_j	C^2_j
1	8	-1.55 (*Reset*) → 0	2.45 (*Reset*) → 0
2	10	0.45	4.45 (*Reset*) → 0
3	9	-0.1 (*Reset*) → 0	3.45 (*Reset*) → 0
4	12	2.45	6.45 (*Reset*) → 0
5	11	3.9	5.45 (*Reset*) → 0
6	15	**9.35** *	9.45 (*Reset*) → 0

* - *Out-of-control signal (upper trend)*

Following the procedures just presented, the target mean rate is calculated as $c = 50 \cdot 0.151 = 7.55$. The Decision Interval H and the parameter F, obtained from Table AII are $H = 8$ and $F = 2$, respectively. Under these circumstances, the reference values (K_1 for the upper trends CUSUM and K_2 for the lower trends CUSUM) are $K_1 = 9.55$ and $K_2 = 5.55$.

The CUSUM values C^1_j and C^2_j were calculated for each sample j and are presented in the third and forth columns of Table II. Notice that the 6th sample C^1_j signals an out-of-control situation ($C^1_6 > H = 8$) that should be studied.

ASSESSING THE REPRODUCIBILITY OF PANELLISTS PERCEPTIONS

The procedure described in this section applies to any sensory attribute evaluated by the panellist during the tasting session. It is assumed that the attribute is represented by a variable expressed in an ordinal scale with at least five levels.

The procedure is particularly effective when the variable refers to the overall quality of the wine. In this situation, each panellist has to rank the goodness of fit between the overall quality of the sample and the expected quality of the wine type. This task is performed using, for example, a five level scale: 1 – "much lower than expected"; 2 – "lower than expected", 3 – "expected," 4 – "better than expected", 5 – "much better than expected". This variable is denoted by Q_k ($k = 1, ..., K$ being the panellist identification).

Periodically and without the knowledge of the panellists, samples are submitted twice to the tasting procedure. The period of time between the tasting of the two samples from the same wine is, at most, one week. This prevents the wine from being significantly altered from the first to the second tasting session. Moreover, as panellists taste between 10 to 20 samples every day, it is not likely that, even if they can identify the wine during the second session, they can remember the rates assigned during the first one.

The reproducibility of each panellist is assessed by the average value of the replication ranges, $\overline{R}_k$ (the range, R_k is the absolute value of the difference between the Q_k values obtained in the original and the replicated session). Based on $\overline{R}_k$, a reproducibility index RI_k was developed:

$$RI_k = 1 - \sqrt{\frac{\overline{R}_k}{E(R)}} \qquad \text{if } \overline{R}_k \leq E(R) \tag{2}$$

$$RI_k = 0 \qquad \text{if } \overline{R}_k > E(R), \tag{3}$$

where $E(R)$ represents the expected value of the range obtained with Q_k values drawn from a uniform discrete distribution. This is equivalent to admit that the panellist rates the samples randomly. Notice that if Q_{max} represents the greatest value of the integer scale used to rank Q_k and Q_{min} the minimum value of the same scale, then

$$E(R) = \frac{(Q_{max} - Q_{min})}{2} \tag{4}$$

Using a five level scale, $Q_{max} = 5$, $Q_{min} = 1$, $E(R) = 2$ and expressions 2 and 3 are as follows:

$$RI_k = 1 - \sqrt{\frac{\overline{R}_k}{2}} \qquad \text{if } \overline{R}_k \leq 2 \tag{5}$$

$$RI_k = 0 \qquad \text{if } \overline{R}_k > 2. \tag{6}$$

The index RI_k varies between 0 and 1. Values of RI_k close to 1 means that the panellist k can reproduce very accurately his perceptions. If the panellist reproduces exactly all the judgements $\overline{R}_k$ is 0 and $RI_k = 1$.
However, if a panellist rates the wines with the same mark in a systematic way (independently of their quality), the value of $\overline{R}_k$ will be zero. In other words,

a value of RI_k close to 1 can also mean that the panellist adopted a "defensive" strategy. This behaviour can be evaluated through another index:

$$DI_k = 1 - \left| \frac{STD_{TP} - STD_k}{STD_{TP}} \right| \quad \text{if } STD_k \leq 2 \cdot STD_{TP} \tag{7}$$

$$DI_k = 0 \quad \text{if } STD_k > 2 \cdot STD_{TP} \tag{8}$$

where:

DI_k: defensive strategy index for panellist k (varying between 0 and 1)

STD_{TP}: standard deviation of the tasting panel

STD_k: standard deviation of panellist k

The standard deviation of the tasting panel, STD_{TP}, reflects the dispersion of the marks assigned by all the TP members to all the wine samples tasted during a particular time period (the replica results must be excluded). This period should be large enough to ensure that the quality of the wine samples submitted to the TP encompasses the full range of the scale. If panellist k uses the scale in the same way as the TP, DI_k takes the value 1. If he adopts a "defensive" strategy it is expected that STD_k (the standard deviation of panellist k) is lower than STD_{TP}. On the other hand, if the panellist uses the scale in a excessive way, STD_k is higher than STD_{TP}. In both situations the value of DI_k is lower than 1, assuming the value 0 when the panellist assigns the same mark to all the wines (or when the extreme values of the scale are excessively used: $STD_k > 2 \cdot STD_{TP}$).

Finally, a reliability index that combines RI_k and DI_k should be computed. This index, I_k, measures the global reproducibility performance of each panellist and is computed as:

$$I_k = RI_k \cdot DI_k . \tag{9}$$

This reliability index can also take values between 0 and 1.

It is suggested that indexes I_k (and hence RI_k and DI_k) are updated continually, if possible each day, using a moving sample. The sample should be drawn using a time period that encompasses more than 20 replica values and a minimum of 200 different wine samples.

Figure 1 presents an example of the evolution of the reliability indexes for 3 tasters over 46 working days (representing a period of about 3 months). This example refers to the evaluation of the overall quality the wine (using a 1 to 5 five scale). The average ranges and the standard deviations are calculated

for the 50 working days preceding the date considered. The number of replicas and wines tasted by each panellist during that period are about 50 and 500, respectively.

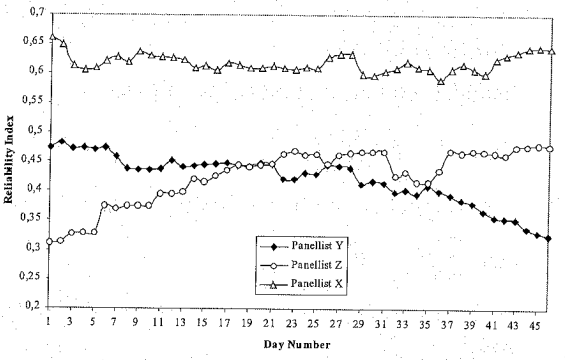

Fig. 1 – Evolution of the Reliability Index for 3 panellists during a 46-day period
Evolução do Índice de Fiabilidade para 3 provadores durante um período de 46 dias

The indexes I_k are currently under testing and will be fully implemented soon.

A NEW PROCEDURE FOR ESTABLISHING THE OVERALL DECISIONS OF TASTING PANELS

The final approval or rejection of a wine sample usually depends on the opinion of the majority of the panellists. For example, if the tasting panel is composed of five members the approval of a sample must be the result of at least three "ones".

Based on the I_k index, a new procedure to establish the overall decisions of tasting panels is proposed (remember that I_k measures the reliability of panellist k). The final decisions ("accepted" or "not accepted") are derived from the weighted sum of the individual decision values assigned by the panellists. Hence, each final decision of a tasting panel composed by K tasters can be expressed as a number, using the formulas:

$$FD_A = \sum_k W_k \cdot DA_k \tag{10}$$

$$FD_R = \sum_k W_k \cdot DR_k \tag{11}$$

where:

FD_A: final decision variable for "approvals"

FD_R: final decision variable for "rejections"

DA_k = 1 if panellist k approves the wine; DA_k = 0 if panellist k rejects the wine

DR_k = -1 if panellist k rejects the wine; DR_k = 0 if panellist k approves the wine

W_k: panellist weight, computed as follows

$$W_k = I_k \cdot K / \sum I_k$$

(the weights W_k are the reliability indexes I_k adjusted in a way that $\sum I_k = K$.

If all the panellists have the same reliability their individual weight is 1).

For a sample to be approved or rejected the value of the final decision variable, $FD_{A,R}$, can not be lower (in the case of FD_A), or higher (in the case of FD_R), than a fixed target. The proposed target value is $|FD_{A,R}| \geq (k+1)/2$: if FD_A is lower than or equal to $(k+1)/2$ and if FD_R is higher than or equal to $-(k+1)/2$ the tasting procedure must be repeated. This strategy does not require an odd number of panellists and takes into account their individual actual performance.

Resumo

A análise sensorial constitui uma parte fundamental no processo de avaliação dos vinhos. No entanto, e apesar da longa tradição que é reconhecida à avaliação sensorial, os métodos existentes para o seu controlo são limitados e insuficientes. Neste artigo descrevem-se alguns dos métodos que foram desenvolvidos durante o processo de acreditação da Câmara de Provadores do Instituto do Vinho do Porto.

Résumé

L'analyse sensoriel est fondamentale dans le processus d'évaluation de vins. Et pourtant, bien qu'elle soit de longue date traditionnellement reconnue, les méthodes existant a son contrôle sont insuffisantes et limitées. Dans cet article on décrit quelques méthodes qui ont étés développés pendant la phase d'accréditation de la Chambre de Dégustateurs de l'Institut du Vin de Porto.

REFERENCES

Agresti A., 1990. *Categorical Data Analysis.* John Wiley & Sons, New York;

Belsley D., 1991. *Conditioning Diagnostics – Collinearity and Weak Data in Regression.* John Wiley & Sons, New York;

Cardinal M., Cornet J., Qannari A., Qannari M., 1994. Performances d'un Groupe d'Évaluation Sensorielle: Exemples de Traitements Statistiques des Données. *Science des Aliments*, **14**, 251-263;

Freund R.J., Wilson W.J., 1998. *Regression Analysis: Statistical Modelling of a Response Variable.* Academic Press, London;

Harrison M.W., Kenneth S.S., Godfrey A.B., 1986. *Modern Methods for Quality Control and Improvement.* John Wiley & Sons, New York;

Hosmer, D. Jr., Lemeshow S., 1989. *Applied Logistic Regression.* John Wiley & Sons, New York;

Lima L.B., Belchior A.P., Estabrook G.F., 1988. Uniformity and Constancy of Wine Tasters Evaluating the Same Wines on Two Different Occasions. *Ciência Tec. Vitiv.*, **7**, 73-85;

Mangan P., 1992. Performance Assessment of Sensory Panelists. *Journal of Sensory Studies*, **7**, 229-252;

Sinesio F., Risvik E., Rodbotten M., 1990. Evaluation of Panelist Performance in Descriptive Profiling of Rancid Sausages: a Multivariate Study. *Journal of Sensory Studies*, **5**, 33-52;

Van Zeller A.L., 1984. Uma Ficha de Exame Organoléptico para Vinhos do Porto. *Ciência Tec. Vitiv.*, **3**, 43-52.

Appendix A – Tables For CUSUM Tabulations

TABLE AI

CUSUM schemes for number of nonconformities (adapted from Harrison et al., 1986, considering scheme C2)

Esquemas CUSUM para número de não conformidades

Target mean rate	*Decision Interval*	*Parameter F*
c	*H*	
0.25	3.0	0.25
0.32	4.0	0.18
0.40	3.0	0.60
0.50	2.0	1.00
0.64	2.0	1.36
0.80	3.5	0.70
1.00	5.0	0.50
1.25	5.0	0.75
1.60	4.0	1.40
2.00	5.0	1.00
2.50	5.0	1.50
3.20	5.0	1.80
4.00	6.0	2.00
5.00	7.0	2.00

TABLE AII

CUSUM schemes for number of nonconforming units (adapted from Harrison et al., 1986, considering scheme C2)

Esquemas CUSUM para número de unidades não conformes

Sample size	*Target mean proportion (p)*	*Decision Interval*	*Parameter F*
(N)		*H*	
20	0.1	3	2
20	0.2	7	1
20	0.3	5	2
25	0.1	4	1.5
25	0.2	5	2
25	0.3	8	1.5
35	0.1	6	1.5
35	0.2	7	2
35	0.3	10	1.5
50	0.1	6	2
50	0.2	10	2
50	0.3	11	2
80	0.1	8	6
80	0.2	13	2
80	0.3	12	3